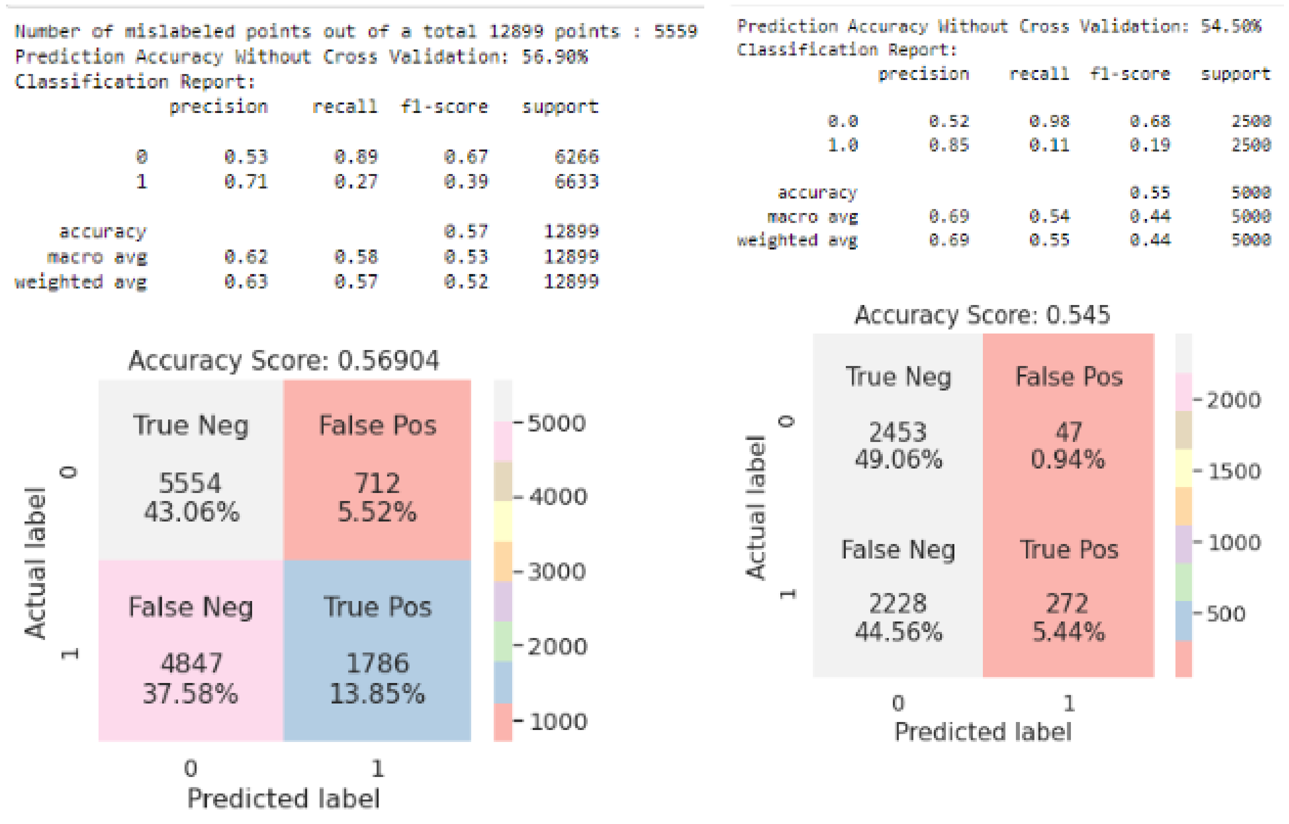
ASSIGNMENT 4

Zeliha Erim 151044065

In this homework, cardio disease dataset used. Its features are these: age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, cardio. There is 11 dependent feature and done independent feature (cardio). Firstly, I have explored data frame, dropped outliers, dropped nan or duplicated rows and visualized features. Two features are skewed so I converted weight and height feature distribution to gaussian distribution using box_cox converter. After this section, I have used five thousand random selected data because training and running algorithms on whole dataset takes too long time, I have written my gaussian Naïve Bayes algorithm and compared its result with python library algorithm result. The results are here:



```
Number of mislabeled points out of a total 12899 points : 5559
Prediction Accuracy Without Cross Validation: 56.90%
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.89      0.67      6266
           1       0.71      0.27      0.39      6633

    accuracy                           0.57     12899
   macro avg       0.62      0.58      0.53     12899
weighted avg       0.63      0.57      0.52     12899
```

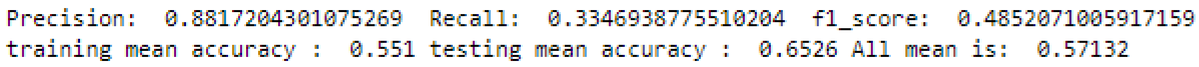
```
Prediction Accuracy Without Cross Validation: 54.50%
Classification Report:
              precision    recall  f1-score   support

         0.0       0.52      0.98      0.68      2500
         1.0       0.85      0.11      0.19      2500

    accuracy                           0.55      5000
   macro avg       0.69      0.54      0.44      5000
weighted avg       0.69      0.55      0.44      5000
```

Figure 1 Python Naive Bayes vs My Naive Bayes

Then I written 10-fold cross validation and details are in notebook however I will place here mean accuracy results:

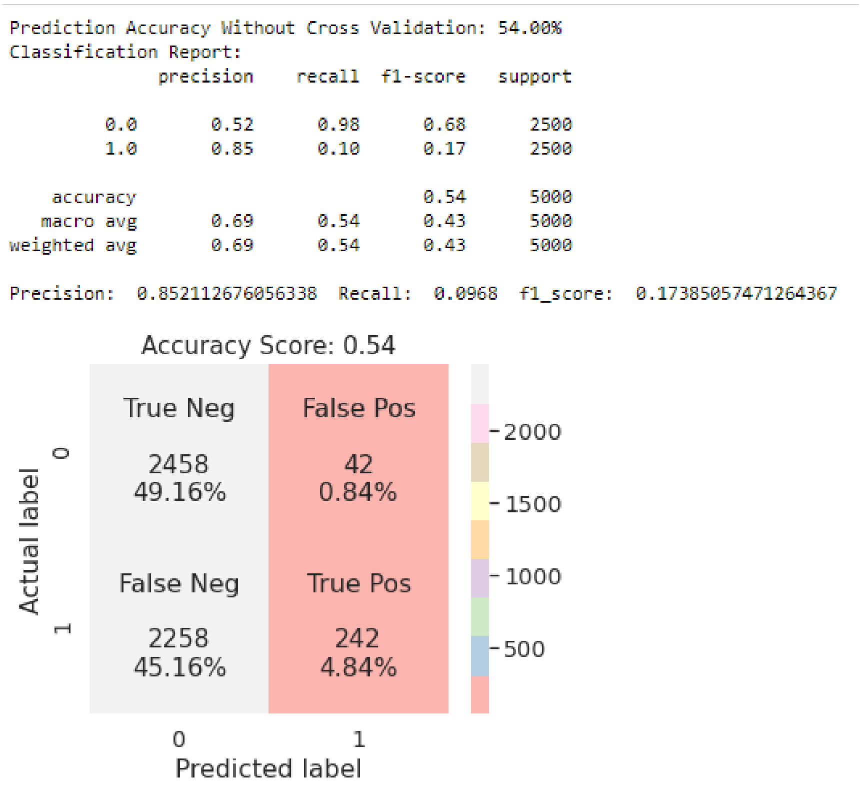


```
Precision:  0.8817204301075269  Recall:  0.3346938775510204  f1_score:  0.4852071005917159
training mean accuracy :  0.551 testing mean accuracy :  0.6526 All mean is:  0.57132
```

Then, to select features, recursive feature selection and Chi square feature selection are used. The first one is wrapper method and the second is filter method. Recursive feature selection is an ensemble model, that is, it uses another algorithm in it for selection features. I have decided this algorithm by visualizing three different machine learning algorithm. These are logistic regression, random forest and XGBRegressor model. After visualizing feature importance scores, the best scores showed by random forest algorithm, thus I used it. I coded feature selection algorithm as backward elimination. In this technique, all features are taken into account, then maximum scored feature selected and deleted, then the rest features taken into account and again maximum scored feature selected and deleted and so on. This will be iterated four times because I want to select four feature. Decided number of features is found out by looking graphs. The other feature elimination algorithm is Chi square. I have divided dataset as zero and one labelled in terms of target variable. Then, calculated expected values to calculated sum of squared difference. The formula is:

$\sum$ (Observe value-Expected value)^2 / Expected value.

Degrees of freedom=(column-1) * (row-1), significant level =0.05 tabular value taken from Chi square table looking significant level and Degrees of freedom values. Then compare results. If X^2 calculated > X^2 tabular then Null Hypothesis is rejected, alternate hypothesis is accepted. Else vice versa case applied. I have used this formula for features and I selected five feature. After that, calculated Naïve Bayes accuracy.



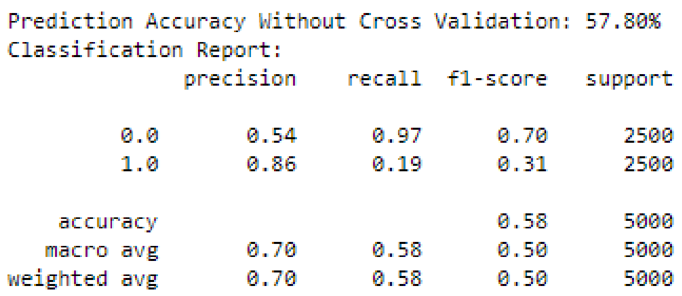
```
Prediction Accuracy Without Cross Validation: 54.00%
Classification Report:
              precision    recall  f1-score   support

         0.0       0.52      0.98      0.68      2500
         1.0       0.85      0.10      0.17      2500

    accuracy                           0.54      5000
   macro avg       0.69      0.54      0.43      5000
weighted avg       0.69      0.54      0.43      5000

Precision:  0.852112676056338  Recall:  0.0968  f1_score:  0.17385057471264367
```

Figure 2 Chi Squared Result



```
Prediction Accuracy Without Cross Validation: 57.80%
Classification Report:
              precision    recall  f1-score   support

         0.0       0.54      0.97      0.70      2500
         1.0       0.86      0.19      0.31      2500

    accuracy                           0.58      5000
   macro avg       0.70      0.58      0.50      5000
weighted avg       0.70      0.58      0.50      5000
```

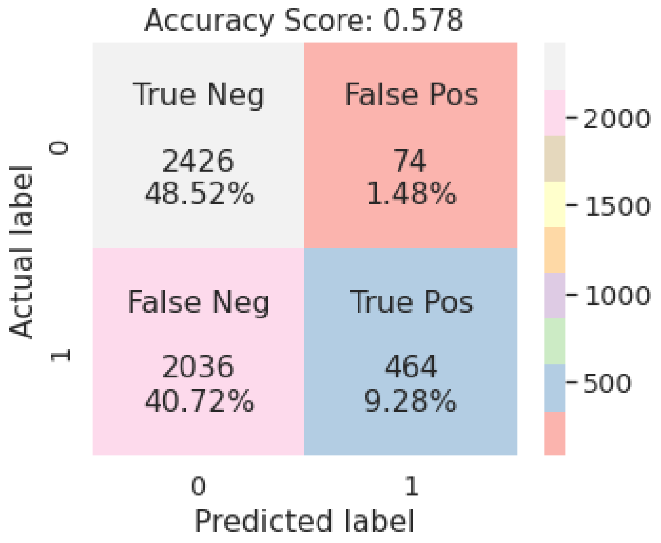
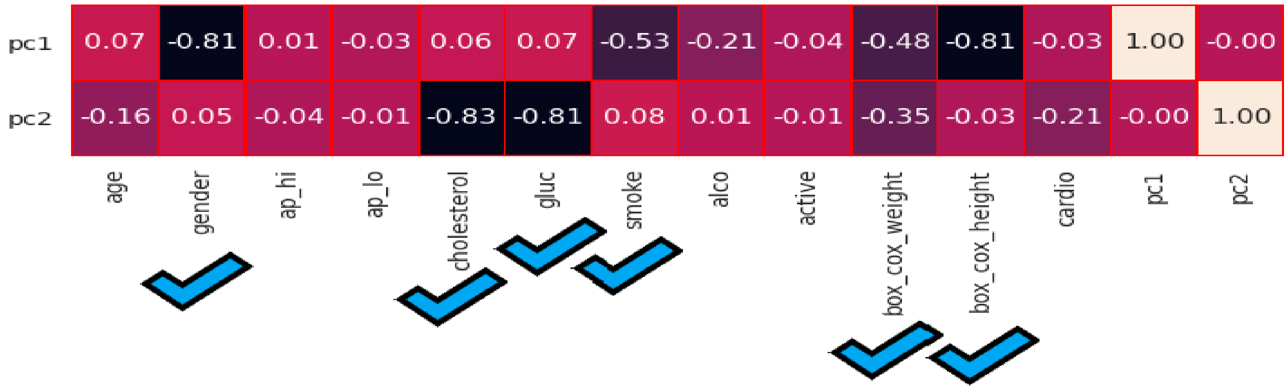
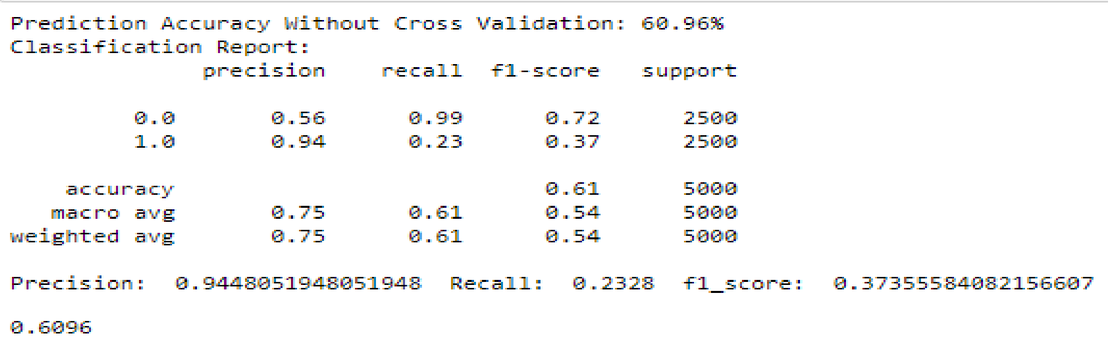


Figure 3Backward Elimination Result

Now, here comes to feature selection using PCA and LDA then comparing their results. In PCA I have used heatmap, looked correlations between features and selected most correlated features. Threshold is 0.3 because some features are changing significantly in this border. I have selected 6 correlated feature and reduced them to 2 features. Then I created again heatmap with these 2 PCA feature and other feature looked correlation ratio. Ratios are high thus algorithm applied very well, also looked PCA features correlation in another heatmap (2*2), their correlation ratio is extremely low, so created features have low cohesion. Before applying PCA, I have used standard scaler for feature behaviors to become same effect on target variable. Then I have used Naïve Bayes algorithm and calculated scores and accuracy like every calculation in this homework. I coded a function to calculate f1 scores, accuracy and recall scores. In LDA target values predicted by Naïve Bayes and LDA algorithm is same. F1 scores are in pictures and notebook notebook.





```
Prediction Accuracy Without Cross Validation: 60.96%
Classification Report:
              precision    recall  f1-score   support

         0.0       0.56      0.99      0.72      2500
         1.0       0.94      0.23      0.37      2500

    accuracy                           0.61      5000
   macro avg       0.75      0.61      0.54      5000
weighted avg       0.75      0.61      0.54      5000

Precision:  0.9448051948051948  Recall:  0.2328  f1_score:  0.37355584082156607

0.6096
```

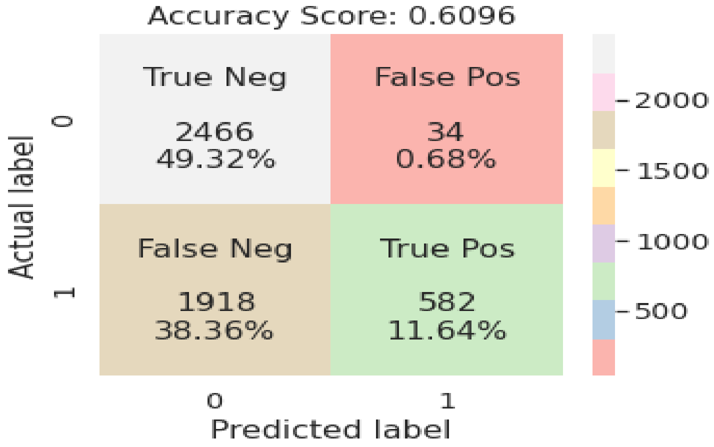
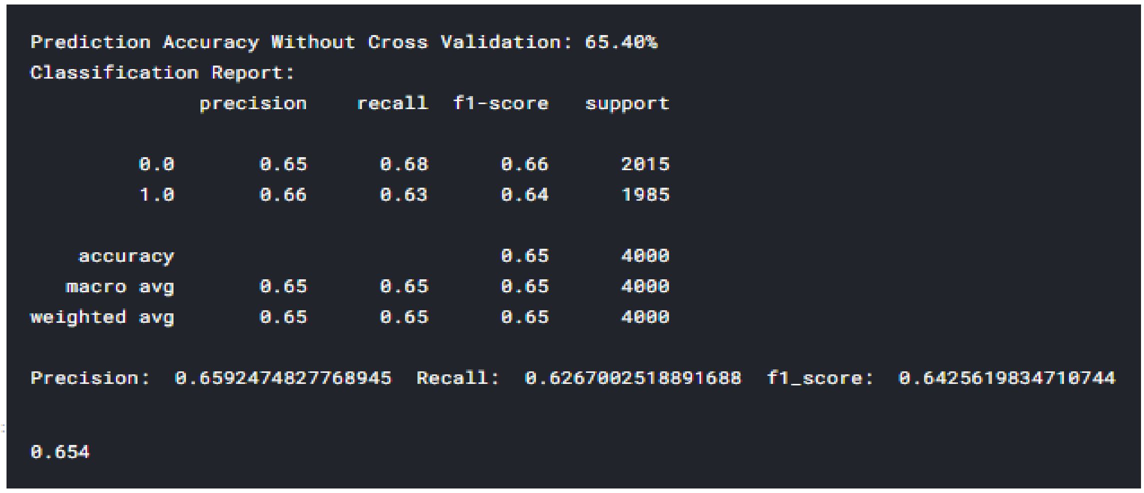


Figure 4 - PCA Confusion Matrix



```
Prediction Accuracy Without Cross Validation: 65.40%
Classification Report:
              precision    recall  f1-score   support

         0.0       0.65      0.68      0.66      2015
         1.0       0.66      0.63      0.64      1985

    accuracy                           0.65      4000
   macro avg       0.65      0.65      0.65      4000
weighted avg       0.65      0.65      0.65      4000

Precision:  0.6592474827768945  Recall:  0.6267002518891688  f1_score:  0.6425619834710744

0.654
```

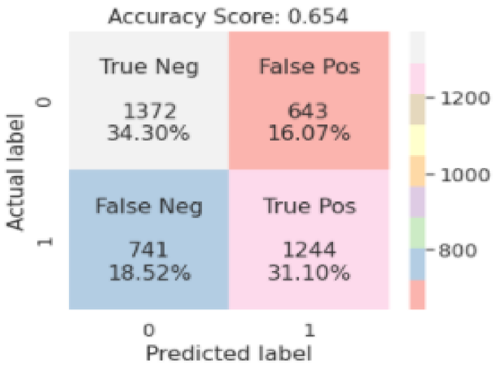


Figure 5 LDA Confusion Matrix Result